

The formal-logical characterisation of lies, deception, and associated notions

Toni Heidenreich

King's College London

Abstract. Defining various dishonest notions in a formal way is a key step to enable intelligent agents to act in untrustworthy environments. This review evaluates the literature for this topic by looking at formal definitions based on modal logic as well as other formal approaches. Criteria from philosophical groundwork is used to assess the definitions for correctness and completeness. The key contribution of this review is to show that only a few definitions fully comply with this gold standard and to point out the missing steps towards a successful application of these definitions in an actual agent environment.

1 Introduction

A formal-logical characterisation of deception and associated notions can be approached with different methods. This review considers two principal methods: formal definitions based on pure modal logic as well as definitions based on agent architecture and communication. We will evaluate how the various formal definitions comply with the philosophical foundations of this topic. Moreover, the respective advantages and disadvantages of these approaches are extracted and used to determine open issues for further research. Putting it most simply, two questions will be answered: Are the existing formal-logical definitions correct given the philosophical background? And what are the problems needing further attention?

Deception has been part of Computer Science since its very beginning. The famous "Imitation Game" proposed by Alan Turing to test whether machines can think [58] explicitly asked for a machine being able to deceive the interrogator. In more practical terms the topic has not been studied until recently when computers became more connected and the concept of agents and multi-agent systems emerged. In 2003, the *AgentLink* community identified trust in agent systems as one of the key challenges for further research [30]. This included user confidence in trustworthiness of machines as well as trust in norms and social rules among agents. As Jones and Firozabadi pointed out [23], trust in this context means trust in the reliability to tell the truth as opposed to other notions like trust in the abilities of others. Dishonesties in this domain have been cause for concern, especially in such application areas as automatic trading agents (e.g. [54]), automatic negotiations (e.g. [43]) or trust in the security of other computer systems (e.g. [56,3]). In particular, in open multi-agent systems where agents are free to leave or enter the system, trust cannot be taken for granted as, e.g., argued in [57].

Since it seems to be established that trust is an important topic in computer science and multi-agent systems in particular, a lot of authors defined trust in a logical-formal way [13,21,22]. This review, however, will focus on the opposite: definitions of not telling the truth, of lying, being deceived or related notions of dishonesty. Although being closely connected, remarkably little work has been done to define these concepts formally, as indicated by many of the authors in the field [35,59,48,36,40]. Summing up, it can be shown that trust is an important topic which has been formally defined by many authors whereas the associated negative notions evolving around lies and deception received comparatively little attention.

In the following chapter 2, we will look at the philosophical work to identify the necessary criteria that definitions of lying and deception need to satisfy in order to be accepted. Chapter 3 will then review some of the existing definitions based on modal logic. Chapter 4 reviews some formal definitions based on other approaches. Finally, chapter 5 will summarise and compare in order to determine open research topics of the area.

2 Philosophical foundation

Many of the papers introduced later base their work on philosophical definitions of lying and deception. Even more important, philosophy has discussed dishonesties for far longer than computer scientists and logicians. Thereby it enables us to identify conditions and criteria that need to be fulfilled by a definition to capture the underlying concepts correctly.

Unfortunately, no definition is universally accepted as different authors indicated [32,44]. Almost every definition can be disputed to encompass either not all or too many cases. Moreover, the boundary between lies and deception is not unambiguous [1]. As recent as 2010, it is claimed that philosophers argue about the right way to define these notions [15]. Despite the ambiguity, we will try to find the conditions supported by the majority of philosophers. In the first section this will be done for the notion of lying, in the second section for deception and the third section will cover other similar notions defined by philosophers. To illustrate the various concepts, the running example of an estate agent guiding a customer through a flat will be used. In this scenario, lying and deceiving may occur very naturally making the example easier to understand.

Lying We will identify five conditions necessary to describe what constitutes a lie. These conditions are based on those expatiated in the Stanford Encyclopedia of Philosophy [32].

Starting off with a dictionary definition, the Cambridge Dictionary says lying is "to say or write something which is not true in order to deceive someone" [6]. This definition includes the first condition, the *statement condition*, meaning that a lie is only a lie if a spoken statement or utterance occurred (defined e.g. by [10]). In the example, the estate agent is therefore lying when he says "I've been in the business for more

than 10 years”, although he knows he just started. He is, however, not lying when he behaves and dresses himself as if he is an expert. Some philosophers argue against this condition [60] maintaining a broader view that even withholding helpful information should be considered a lie.

What the dictionary definition is probably doing wrong, is to assert that the utterance must be false. This is best illustrated with the case where the estate agent honestly believes that the flat has been renovated recently although it has not. According to the definition, he would be lying by telling the customer that the flat has been renovated recently, although it is his honest believe. To exclude this dilemma, most philosophers introduce the *believe-false condition*, where the liar must believe the proposition he is saying to be false, independent of the actual truth-value. This is reflected, for example, in the definition of J. Kupfer that “a person lies when he asserts something to another which he believes to be false with the intention of getting the other to believe it to be true” [28]. Similar definitions have been given by I. Primoratz [38] or B. Williams [62]. Some like Chisholm and Feehan [10] argue that a *not-believe-true condition* is also enough which includes the case where the liar has no believe at all about what he is saying.

Kupfer’s definition is also clarifying the role of the listener. First of all there must be a listener, usually called the *addressee condition* [32]. Thereby cases of telling some wrong believe to an empty room or being eavesdropped telling a wrong belief are not lies per se. Furthermore, the definition highlights the *intent-to-deceive condition* which has also been described by J. Mahon [33]. Our estate agent joking about the value of the property by telling an exorbitant price is not lying, because he has no intent to deceive the customer with this statement. Even if some like R. Sorenson [55] or T. Carson [8] argue against this condition in special cases, general adaptations should include this condition to rule out fakes, jokes or play-acts [15].

The fifth condition is the demand that a definition should in no case include the *success condition*, meaning that the intended deception was successful [32]. Imagine the case when the estate agent lies about the quality of the parquet and the customer happens to be a skilled carpenter being able to judge the quality correctly. In this occasion he would still think that the estate agent lied, so the success (or failure) of the attempted deception makes no difference to the fact that a lie occurred.

It is interesting to note that none of the mentioned conditions state that a lie is morally wrong [32]. As reasoned by Kemp and Sullivan this is because morality is “a synthetic judgement and not an analytic one” [27]. There are also other possible conditions for lying which are, for example, defined by C. Sakama in [44]. Since he also gives formal definitions, we won’t use his conditions to ensure that the standard is not biased towards his definition.

This leaves us with the five conditions from the Stanford Encyclopedia of Philosophy that a lie is only correctly defined if it contains the statement condition, the believe-false-condition, the addressee condition, the intent-to-deceive condition and not the success-condition or any other restrictions.

Deceiving The Cambridge Dictionary defines deceiving as "persuade someone that something false is the truth" [6]. This definition shows us that deception is concerned with the effect on the listener or receiver of the message as opposed to lying which is focused on the dishonest behaviour of the speaker. For this reason, three changes need to be made to the list of necessary conditions.

As already indicated, the first change is to include the *success condition*. When the estate agent lied about the parquet, he did not manage to deceive the carpenter. Deception would have only occurred if the customer had believed the lie. As argued by Chisholm and Feehan in [10], the proposition the listener believes after the deception does not necessarily need to be a new belief. He could also be deceived by maintaining a belief or even by being prevented from acquiring some belief (in this case the success is that he continues to believe that the proposition is not true).

The second change is to exclude the *statement condition* as reasoned by J. Mahon [32] or L. Linsky [29]. Coming back to a previous example, a knowledgeable appearance of the estate agent does not constitute a lie. But if this is intentional, the customer could be deceived about his expertise without any statement being made. Instead, a new criterion called *evidence condition* is added. It says that the deceiving person must provide some form of evidence which is the reason for the listener to conclude the wrong proposition [19,4,31]. This ensures the agency of the deceiving person. For example, if a friend of the customer told him that the estate agent has a high level of expertise and the agent still wears the potentially misleading outfit and the belief of the customer is solely based on the friend, then the estate agent has definitely not deceived the customer. In another example, the estate agent might even tell the truth by saying that the parquet is shining brightly with the intention that the customer concludes a high quality. Still, this is a form of deception as evidence intentionally makes the customer believe something that the estate agent does not believe.

All other conditions still hold for deception: the *believe-false* or its weaker form *not-believe-true condition*, the *addressee condition* and the *intent-to-deceive condition* (e.g. [31]). Some like J. Adler argue that deception does not need to be intentionally [1]. But as most of the other authors disagree with this position we will still use it as a necessary condition.

As a result, we know that a definition of deception needs to fulfil the believe-false condition, the addressee condition, the intent-to-deceive condition, the evidence condition and the success condition. It must not include the statement condition or other restricting notions.

Other notions The definitions of dishonest behaviours given by philosophers are not limited to lies and deception, although being the main focus. Other notions which have been considered are fraud, bullshit, withholding information or half-truths. They became necessary as the available definitions of lying and deceiving did not include all kinds of possible dishonesties.

M. Simmons defined in [52] the notion of *fraud*. His definition essentially includes a lie which is believed by the listener and therefore also becomes a successful deception, combined with requirement that the victim uses the acquired information and makes loss of money or property as a result. This definition shall not be used as a guideline in this review but is given here because one of the formal definitions refers to it.

A more relevant notion is *bullshit* as it was denoted by H. Frankfurt in [18]. Bullshit covers similar cases to lying with the difference that the speaker does not follow the believe-false condition and instead neither believes the statement to be true nor does he believe it to be false. The estate agent bullshits, for instance, when he highlights the satisfaction of the previous tenants even though he did not know them.

Another important notion is *withholding information* as e.g. mentioned by T. Carson in [8]. It is the failure to offer information that would help to acquire true beliefs or correct false beliefs as long as this result is intentional. For example, not telling that the previous tenant moved out because of noisy neighbours is an obvious example of withholding information. In this case, the conditions for lies can be adopted as well, with the difference that the statement condition is replaced by the explicit non-statement that would have helped to change the belief of the listener. Carson also mentions *half-truths* in the same paper as a special case of deception including a true statement as evidence.

Although a number of other notions have been defined, mostly bullshit and withholding information will be relevant for the formal definitions.

Altogether, this chapter provided a basic overview of how philosophers define lies, deception and similar concepts. For lies and deception we were able to extract a concrete list of conditions allowing us to check whether formal definitions comply with this standard.

3 Modal logic definitions

This chapter will look at definitions of lies and deception in modal logic. Using logic to describe these concepts is an obvious decision as it provides a very general expressiveness which is well understood and not restrained to an application area. However, as standard propositional logic is not enough to capture all the subtle notions, almost all definitions make use of some kind of modal logic. By introducing additional modal operators, this kind of logic allows for quantifications not possible with simple true/false values (cf. Stanford Encyclopedia of Philosophy [20]). Different authors use different sets of operators; for simplicity we will introduce the five most common and explain less common operators later when they are applied:

- $\mathcal{B}_i p$ denotes that agent i *believes* the fact p . This does not say anything about the underlying truth of p and just indicates the beliefs of agent i . This operator and the following two often occur together in a system called *BIC* which was e.g. specified by M. Colombetti [11].

- $\mathcal{I}_i p$ denotes that agent i has the *intention* to make p true. For example, by writing $\mathcal{I}_i \mathcal{B}_j p$ we can specify that the estate agent i has the intention that the customer j believes that the flat is renovated ($=p$).
- $\mathcal{C}_{ij} p$ denotes that agent i *communicates* the fact p to agent j . This includes any kind of spoken communication or utterance.
- $\mathcal{O} p$ denotes that it *ought to be* p , where p is any proposition. The operator is e.g. described by B. Chellas [9] or A. Jones [24] and means that in known environments it is ideally the case that p . For example in the flat sale situation, it ought to be the case that the estate agent is authorised to sell the flat.
- $\mathcal{E}_i p$ denotes that an agent i *brings about* that p and was introduced by I. Pörn [39]. It assigns the agency of i to the fact p , in the sense that agent i is the decisive factor that p occurred or became true. Then a friend of the customer told him that the estate agent is an expert in his field, this friend brought about that the customer believes in the expertise.

Authors of papers using these operators usually provide proofs or refer to others who proofed that the logical system is coherent and fulfils a number of desired properties. We won't go into details at this point as interested readers may refer to the original papers.

In the following, we will start to look at formal definitions of lying before continuing with bullshit, deception and other notions. Each considered paper is evaluated using the conditions selected in the last chapter. Within each group a chronological order will be maintained if reasonable. Thus, developments and improvements over time can be seen.

Lying

B. O'Neill (2003) [35]. The first definition we will examine is one given by B. O'Neill. In his paper, he defines and derives several properties of the modal operator \mathcal{C} for communication and subsequently defines lying with the formula in equation 1. Starting from his previously derived properties of communication he then shows that lying is a subset of all situations which satisfy equation 2.

$$\mathcal{C}_{ij} p \wedge \mathcal{B}_i \neg p \tag{1}$$

$$\mathcal{I}_i \mathcal{B}_j \mathcal{B}_i p \wedge \mathcal{B}_i \neg \mathcal{B}_i p \tag{2}$$

In an example, this might mean that the estate agent i is lying when he tells customer j that the flat has been renovated ($\mathcal{C}_{ij} p$) while believing this is not true ($\mathcal{B}_i \neg p$). As a matter of some rules governing communication it implies that he intends the customer to believe that he believes the flat has been renovated ($\mathcal{I}_i \mathcal{B}_j \mathcal{B}_i p$) and that he does not believe that the renovation is part of his belief ($\mathcal{B}_i \neg \mathcal{B}_i p$).

It can be seen very easily that the statement condition ($\mathcal{C}_{ij} p$), the addressee condition (j) and the believe-false condition ($\mathcal{B}_i \neg p$) are fulfilled. The definition also contains no notion of success and some intention. The problem is, however, that this intention is not exactly the desired intent-to-deceive condition as agent i is not intending that j actually believes

p , but rather that j accepts that i is telling his true belief ($\mathcal{I}_i\mathcal{B}_j\mathcal{B}_ip$). By adding this extra level of abstraction, O'Neill fails to meet this condition.

On the other hand he shows very well that besides the believe-false condition, the weaker not-believe true condition can constitute some form of dishonesty which he calls 'talking though one's hat'. The according formal definition just replaces $\mathcal{B}_i\neg p$ with $\neg\mathcal{B}_ip$ in equation 1.

As a result, we can say that O'Neill gives a well-thought-of definition with a small inconsistency and manages to relate different levels of lying in the formal definition.

M. Caminada (2009) [7]. This paper of Caminada mainly focuses on the difference between lying and bullshit. In this context he defines lying as given in equation 3. In parts, this definition is very similar to one given by A. Tzouvaras 11 years earlier [59], but is more clarified by leaving out unnecessary parts.

$$\mathcal{C}_ip \wedge \mathcal{B}_i\neg p \quad (3)$$

His definition complies with the statement condition (\mathcal{C}_ip), the believe-false condition ($\mathcal{B}_i\neg p$) and it contains no notion of success. However, it suffers from the problem of not including anything concerned with the listener, so that neither the addressee condition nor the intent-to-deceive conditions are satisfied.

Despite the lacking expressiveness of this definition he claims that the definition of lying is settled and well-defined. He mentions, however, that the intent-to-deceive condition could be added but argues that this easier approach is sufficient. Nevertheless, the formal definition as it was given should be rejected for the named reasons.

C. Sakama et al. (2010) [48]. In this paper the authors try to formally define lying, as well as bullshit and deception. The other definitions besides lying will be given later for clarity reasons.

Using a similar logical framework as all aforementioned authors, their definition of lying in equation 4 is the first satisfying all the conditions.

$$\mathcal{C}_{ij}p \wedge \mathcal{B}_i\neg p \wedge \mathcal{I}_i\mathcal{B}_jp \quad (4)$$

It contains a statement ($\mathcal{C}_{ij}p$) to an addressee, the speaker fulfils the believe-false condition ($\mathcal{B}_i\neg p$), the intent-to-deceive condition ($\mathcal{I}_i\mathcal{B}_jp$) and no notion of success is included. The definition is therefore fully compatible with the philosophical criteria.

In addition, the authors give more specialised versions of this definition which include the objective of the liar. They also conclude that lies have to be as weak as possible to deceive the listener as they always introduce some deviation from the truth (or from what the liar believes to be the truth). That later binds him to his lie and makes him less free in what he can say without contradicting himself. This observation is quiet application-oriented and shows more insight than other papers. All in all, this definition shows to be the most comprehensive definition among those which have been reviewed, both in accuracy and profundity.

Bullshit

M. Caminada (2009) [7]. The first and oldest formal definition of bullshit is the one given in Caminada's paper which already included a definition of lying. Since bullshit as a philosophical concept was only defined in 2005 by H. Frankfurt [18] this is just one of two available formal definitions. Bullshit in Caminada's definition in equation 5 highlights the main difference to lying: that the speaker has neither a belief that his statement is true ($\neg\mathcal{B}_ip$) nor that it is false ($\neg\mathcal{B}_i\neg p$).

$$\mathcal{C}_ip \wedge \neg\mathcal{B}_ip \wedge \neg\mathcal{B}_i\neg p \quad (5)$$

Similar to his definition of lying, it also contains a statement (\mathcal{C}_ip), but no addressee and no intention to deceive. In contrast to lying, this is not necessarily a problem as Frankfurt's informal definition does not specifically contain these parts either. Lending the example given by C. Sakama in [48], we can imagine that the estate agent is providing a consulting service and is paid per hour or report length. This might cause him to produce some bullshit according to equation 5 just to earn more money but without any intent that the reader is actually believing what he has written, since it makes no difference as long as he appears knowledgeable.

C. Sakama et al. (2010) [48]. Sakama et al. present bullshit as a weaker form of lying in the already mentioned paper. They actually give exactly the same definition as Caminada (Eq. 5). Furthermore, they produce a definition for 'intentional bullshit' (Eq. 6) including the missing intent-to-deceive condition ($\mathcal{I}_i\mathcal{B}_jp$) and the addressee condition.

$$\mathcal{C}_{ij}p \wedge \neg\mathcal{B}_ip \wedge \neg\mathcal{B}_i\neg p \wedge \mathcal{I}_i\mathcal{B}_jp \quad (6)$$

Relating to their consideration about the impact lying may have on future communication, they conclude that intentional bullshit should be preferred over lying, if possible, as it does not deviate from the true belief as much as a lie. This conclusion seems quite natural as people are more likely to tolerate bullshit than lies as H. Frankfurt pointed out in his work [18].

Deception

B. Firozabadi et al. (1998) [16]. This paper by Firozabadi et al. is focused on verifying trade procedures by excluding fraud. Based on the definition of fraud given by M. Simmons [52] they produce four different possible formal definitions of deception (which is a constituting part of fraud). In contrast to all other authors we looked at so far, they use the modal operators \mathcal{B} for belief, \mathcal{E} for 'brings about' and a derived operator \mathcal{H} for 'attempts to bring about'. The latter was first introduced by F. Santos et al. [50] and has the same meaning as 'bringing about'

something but without the inherent success of the action.

$$\neg \mathcal{B}_i p \wedge \mathcal{E}_i \mathcal{B}_j p \quad (7)$$

$$\neg \mathcal{B}_i p \wedge \mathcal{H}_i \mathcal{B}_j p \quad (8)$$

$$\mathcal{B}_i \neg p \wedge \mathcal{E}_i \mathcal{B}_j p \quad (9)$$

$$\mathcal{B}_i \neg p \wedge \mathcal{H}_i \mathcal{B}_j p \quad (10)$$

In the deception definitions (Eq. 7 to 10) they include the believe-false ($\mathcal{B}_i \neg p$) or the weaker not-believe-true condition ($\neg \mathcal{B}_i p$) and the deception is directed at an addressee (j) without the notion of communication. Both operators, \mathcal{E} and \mathcal{H} , denote the agency of i and do not include any intention. Therefore, the intent-to-deceive condition is not included. Moreover, the \mathcal{H} operator does not even include the success of the deception, that is why equations 8 and 10 need to be rejected. The last criterion asking for evidence is not included as well. In summary, these definitions lack several of the necessary conditions and are not up to the standard given by the philosophical literature.

A. Jones et al. (2001) [23]. In this paper A. Jones and B. Firozabadi (the author of the previous paper) improve on the definition of deception. Again, they use the belief operator \mathcal{B} and the operator \mathcal{E} for 'bringing about' something. Additionally, they use the already introduced operator \mathcal{O} for 'it ought to be that' and another operator $a \Rightarrow_s b$ denoting that a 'counts as' b given the context or institutionalised power of s . Simplified, the operator which was introduced by A. Jones in [25] denotes a consequence which is true under given circumstances. For example, getting a plastic card with one's name on it (a) 'counts as' being a student (b) as long as this is done by an university (s).

Their definition of deception in equation 11 fulfils all conditions except the intent-to-deceive condition.

$$\neg \mathcal{B}_i p \wedge \mathcal{E}_i \mathcal{B}_j \mathcal{E}_i m \wedge (((\mathcal{E}_i m \Rightarrow_s \mathcal{O} p) \wedge \mathcal{B}_j \mathcal{E}_i m) \rightarrow \mathcal{B}_j p) \quad (11)$$

It doesn't contain a statement, but an addressee (j) and it complies with the not-believe-true condition ($\neg \mathcal{B}_i p$). The evidence condition is included ($\mathcal{E}_i \mathcal{B}_j \mathcal{E}_i m$ and $\mathcal{E}_i m \Rightarrow_s \mathcal{O} p$) in a way that i brings about that j believes he brought about the evidence m , while under the current circumstances s this bringing about of evidence m ought to mean that p is true. The success of the deception is included as well, as this allows j to reason that p is true ($\rightarrow \mathcal{B}_j p$). The missing intention to deceive is, however, only a minor problem as this was one of the disputed conditions for deception anyway (as mentioned in chapter 2).

G. Meggle (2000) [34]. Unlike the previous authors, Meggle uses a variant of the \mathcal{BIC} logic to describe deception and associated notions.

$$\mathcal{I}_i \mathcal{B}_j p \wedge \mathcal{B}_i \neg p \wedge \mathcal{C}_{ij} m \wedge \mathcal{B}_i (\mathcal{C}_{ij} m \rightarrow \mathcal{B}_j p) \wedge (\mathcal{C}_{ij} m \rightarrow \mathcal{B}_j p) \quad (12)$$

The definition he gives for successful deception in equation 12 correctly contains the intent-to-deceive ($\mathcal{I}_i \mathcal{B}_j p$) and the believe-false condition

$(\mathcal{B}_i \neg p)$. It provides evidence $(\mathcal{C}_{ij}m)$ which i believes to cause j to acquire the new belief $(\mathcal{B}_i(\mathcal{C}_{ij}m \rightarrow \mathcal{B}_jp))$ and it also covers the success in the way that this evidence indeed causes j to acquire the new belief $(\mathcal{C}_{ij}m \rightarrow \mathcal{B}_jp)$. The problem of this definition is, however, that it explicitly contains a statement condition which should be avoided to account for deception by other means.

Furthermore, Meggle uses his definition to nest multiple levels of deception (being deceived about being deceived and so on) and thinks about implications for actual implementation. These considerations and the good, albeit not perfect, definition contribute to make this a valuable paper.

B. O'Neill (2003) [35]. In the already mentioned paper of B. O'Neill, he also gives a definition of deception as in equation 13. In his paper he actually starts by defining deception, only to show later that his definition of lying is a subset of deception.

$$\mathcal{I}_i \mathcal{B}_jp \wedge \mathcal{B}_i \neg p \wedge \mathcal{B}_jp \quad (13)$$

He correctly leaves out a statement, but has an addressee (j), fulfils the believe-false condition $(\mathcal{B}_i \neg p)$, the intent-to-deceive condition $(\mathcal{I}_i \mathcal{B}_jp)$ and the success condition (\mathcal{B}_jp) . Again, one condition is not fulfilled, as no trace of an evidence is included in his definition.

C. Sakama et al. (2010) [48] In the same paper where they already defined lies and bullshit, Sakama et al. also define deception with equation 14. They try to focus on the speaker's point of view, thereby neglecting some of the necessary conditions.

$$\mathcal{C}_{ij}m \wedge \mathcal{B}_im \wedge \mathcal{I}_i \mathcal{B}_jm \wedge \mathcal{B}_i \mathcal{B}_j(m \wedge \neg \mathcal{B}_j \neg p \rightarrow p) \wedge \mathcal{B}_i \neg \mathcal{B}_j \neg p \wedge \mathcal{B}_i \neg p \wedge \mathcal{I}_i \mathcal{B}_jp \quad (14)$$

The way to interpret this rather long statement is that i communicates some evidence m which he believes himself and intends j to believe it as well. He furthermore thinks that j makes the default conclusion that p holds as well (as long as he does not believe that $\neg p$). Agent i also expects that j does not believe $\neg p$ while he himself does not believe in p with the overall intention the j comes to believe p .

It contains an addressee, the believe-false condition, the intent-to-deceive and some evidence. The problems are that it does not state at any point that the deception was successful and that it introduces unnecessary restrictions by using the statement condition and restricting the evidence to propositions he believes himself (\mathcal{B}_im) .

Even though the definition lacks several of our criteria, the authors do well in concluding that deception is even better than bullshit or lying as it does not even require to deviate from the truth at all.

C. Sakama et al. (2010) [47]. In this paper of C. Sakama and M. Caminada, the authors try a different approach to define deception by enumerating all possible situations which might constitute deception. They base these

definitions on the work of Chisholm and Feehan [10] who differentiated deception by aim, effect and knowledge. Besides the familiar operators of belief (\mathcal{B}), intention (\mathcal{I}), communication (\mathcal{C}) and bringing about something (\mathcal{E}), they additionally use 'let it be the case that' denoted by \mathcal{F} . This operator has a similar meaning to \mathcal{E} , but with the difference that $\mathcal{E}_i p$ denotes that the agency of i changes p , while $\mathcal{F}_i p$ means that the agency of i allows p to continue to be what it was before.

The definitions in equations 15 to 22 can be summarised as telling or not telling something which causes the listener to start believing, continue believing, cease not believing or being prevented from not believing some proposition.

$$\mathcal{B}_i \neg p \wedge \mathcal{C}_{ij} p \rightarrow \mathcal{E}_i \mathcal{B}_j p \quad (15)$$

$$\mathcal{B}_i \neg p \wedge \mathcal{C}_{ij} p \rightarrow \mathcal{F}_i \mathcal{B}_j p \quad (16)$$

$$\mathcal{B}_i \neg p \wedge \mathcal{C}_{ij} p \rightarrow \mathcal{E}_i \neg \mathcal{B}_j \neg p \quad (17)$$

$$\mathcal{B}_i \neg p \wedge \mathcal{C}_{ij} p \rightarrow \mathcal{F}_i \neg \mathcal{B}_j \neg p \quad (18)$$

$$\mathcal{B}_i \neg p \wedge \neg \mathcal{C}_{ij} \neg p \rightarrow \mathcal{E}_i \mathcal{B}_j p \quad (19)$$

$$\mathcal{B}_i \neg p \wedge \neg \mathcal{C}_{ij} \neg p \rightarrow \mathcal{F}_i \mathcal{B}_j p \quad (20)$$

$$\mathcal{B}_i \neg p \wedge \neg \mathcal{C}_{ij} \neg p \rightarrow \mathcal{E}_i \neg \mathcal{B}_j \neg p \quad (21)$$

$$\mathcal{B}_i \neg p \wedge \neg \mathcal{C}_{ij} \neg p \rightarrow \mathcal{F}_i \neg \mathcal{B}_j \neg p \quad (22)$$

The eight definitions exist without or with the intentional part $\mathcal{I}_i \mathcal{B}_j p$. By this they include the intent-to-deceive condition while not ruling out the possibility of unintentional deception. They clearly also have an addressee and a believe-false condition ($\mathcal{B}_i \neg p$). Furthermore all possible forms of success are included. At first it might seem problematic to include the statement condition ($\mathcal{C}_{ij} p$), but by enumerating the same equations with the explicit non-statement ($\neg \mathcal{C}_{ij} \neg p$) they actually show indifference to the statement as required by the philosophical criteria. The only thing that is definitely missing is the evidence condition.

All together, this definition stands out in its achievement to include all the different notions which might be included in deception with the downside that the evidence condition is missing completely.

Other notions

C. Sakama et al. (2010) [47]. Other notions of dishonesty defined in a formal way were only embedded in papers of C. Sakama. One of them is 'withholding information' which is defined in the same way as lying but with the non-statement instead of the statement condition (Eq. 23).

$$\neg \mathcal{C}_{ij} \neg p \wedge \mathcal{B}_i \neg p \wedge \mathcal{I}_i \mathcal{B}_j p \quad (23)$$

This definition can be considered as complete, since the non-statement condition, the addressee condition, the believe-false condition and the intent-to-deceive condition are included without using any success condition.

C. Sakama (2012) [45]. This presentation of Sakama includes a number of his previously mentioned definitions and additionally a definition of half-truths. Interestingly, half-truths have the same definition as his deception definition in [48], given in equation 14. Presumably, he noticed the shortcomings of this definition as deception and relabelled it as half-truth which is indeed more appropriate given, e.g., the dictionary definition that a half-truth is "a statement that is intended to deceive by being only partly true" [6].

Summary

After looking at a number of definitions in modal logic, we can conclude that fulfilling the various philosophical criteria is not at all obvious. For lying, only Sakama's definition in [48] is fully compatible with the criteria. For deception, none of the reviewed papers included a definition which is totally correct. The definitions of Jones [23], Meggle [34] and Sakama [47], however, are the most suitable with only one condition missing in each definition. The latter manages to provide very subtle differences by using multiple statements with the downside of introducing possibly unnecessary complications. On the other hand, the definitions of the minor notions of bullshit and withholding information seem to be quite accurate though not many tried to define these notions.

A problem which definitions based on modal logic have in common, is that they are coupled to this rather complicated form of logic with different operators which are usually not used in the application context of agent design. Therefore, more work is needed to transfer the results to actual agents, possibly by adapting simpler operators. This step might of course damage the precision of the definitions but may help to apply them in practice. Furthermore, research is needed to come up with logically proofed methods to reason about when to use lies and how to recognise them. Sakama's considerations in [48] about the preferred way of using dishonesty by deviating as little as possible from the truth are a first step in this direction. On the other hand, the approaches introduced in the next chapter might be more useful for practical applications after all, as they already embed formal definitions within agent communication and argumentation.

4 Other approaches to formal definitions

This chapter will examine other approaches and methods to define lies, deception and other notions. These are mainly embedded in existing agent architecture and communication systems bringing the advantage of being in the application context already. Nonetheless, these definitions can be written in a formal way to make them relevant to this review and to allow comparisons with the previously presented definitions with modal logic.

We will examine three examples to see how the definition can be embedded at very different levels of abstraction. The first example will

look at a low-level approach operating on the belief base of an agent. The second paper uses a mid-level definition operating on speech acts and thirdly, a high-level definition will be given which is embedded in abstract argumentation frameworks.

Subsequently, we will begin to look at a low-level approach before continuing with higher level definitions.

Low-level approaches An intelligent agent usually has a continuous update cycle of sensing the environment, updating the current beliefs about the world, deciding what to do by using desires and intentions and acting according to a plan which might achieve the current intention (following the idea first introduced by M. Bratman in [5]). The belief base of the agent which is updated in every cycle is the starting point of the low-level approaches. This way they allow for similar flexibility as modal logic while still being embedded in the agent's sense-decide-act cycle.

One example we won't examine in detail is given by F. De Rosi et al. in [12], where they employ a probabilistic model to model the belief base and define lies as a set of conditional probabilities thereupon.

We will look in more detail at a presentation of C. Sakama of 2012 [46] which is based on a previous paper of Sakama et al. [49] where they introduce 'logic programming' as a way to represent the knowledge base of an agent which allows to include disinformation. By disinformation they refer to all possible lies or bullshit given a certain agent's belief set.

The simplified version presented here uses the pair $\langle K, D \rangle$ as the belief base of an agent, such that K is the agent's knowledge and D contains all propositions that count as lies or bullshit according to equation 24.

$$\forall l \in D, K \models \neg l \vee (K \not\models l \wedge K \not\models \neg l) \quad (24)$$

Additionally, the agent has the explicit goal to propose g although it is not in the knowledge base ($K \not\models g$) or to prevent g from being proposed although it is in the knowledge base ($K \models g$). For this purpose, the agent can use an adapted knowledge base $(K \setminus J) \cup I$, where $J \subseteq K$ and $I \subseteq D$. Putting it more simply, the agent can ignore some parts of the knowledge base counteracting his intentions (J) and add some disinformation (I) enabling him to achieve his goal.

Lies, bullshit and withholding information are now defined using these sets J and I (Eq. 25 to 27).

$$\text{Lie, if } I \neq \emptyset \wedge K \models \neg l \text{ for some } l \in I \quad (25)$$

$$\text{Bullshit, if } I \neq \emptyset \wedge K \not\models \neg l \text{ for any } l \in I \quad (26)$$

$$\text{Withholding information, if } I = \emptyset \quad (27)$$

Intuitively, lying is adding some additional information of which at least some is believed to be false. Bullshitting is adding some additional information of which none is believed to be false (nor to be true, as this would have enabled the agent to use the original set K without adding anything). Finally, withholding information is leaving some information out without adding any disinformation.

The statement, addressee and intent-to-deceive conditions are not included explicitly in the definition. However, the framework implicitly includes all three conditions as the whole process is aimed at using the modified believe set to communicate p or prevent p from being communicated to someone with the intent to deceive (since communicating p would not have been possible with the original belief K).

The definition of lying includes the believe-false condition and no notion of success and therefore fulfils all criteria. The same holds for the bullshit definition as it contains the indifferent believe and no notion of success either. Finally, withholding information fulfils all criteria as well, in this case by not communicating essential parts.

In his presentation, Sakama also gives some behavioural rules when to use each of the possibilities, using the preference ordering he already introduced in his modal logic papers of truth over withholding information over bullshit over lies.

Given that he manages to give definitions complying with all necessary criteria and nonetheless being incorporated in an actual agent design, this approach is one that should be taken into account.

In summary, the low-level approaches share the problem of either leaving out necessary conditions or including them only implicitly through the framework they are used in. However, if this implicit inclusion is acknowledged, at least the approach of Sakama in [46] fulfils all criteria and, moreover, adds rules of how to apply the dishonesties in communication.

Mid-level approaches The next level of abstraction is reached when the agents need to communicate with one another in a multi-agent system. The communication can be of various kinds, e.g. deliberation, inquiries, negotiation, persuasion or info-seeking as defined by Walter and Krabbe in [61]. The constituting parts of communication are speech acts each consisting of a performative and the content, where the performative denotes the kind of speech (like requests, promises or assertions, see the classification of J. Searle [51]). These speech acts, which are usually defined in a generally accepted language like FIPA-ACL [17], can be modified to accommodate for lies and deception.

The paper of E. Sklar et al. [53], we will look at in more detail, uses a speech-act based approach to integrate lies in agent communication. The main problem they face is the need to use the existing agent communication languages which are designed for truthful communication as pointed out by Parsons and Wooldridge in [37]. These existing languages allow a finite list of defined performatives for speech act, that is why no new performative 'lying' could be added. As Sklar et al. argue, this would not make sense anyway as the performatives are public and an agent who is publicly announcing that he is lying cannot really lie.

What they are doing instead is to alter the existing pre- and post-conditions of the performative 'assert' to allow an agent to make false assertions (i.e. to lie). The original definition of 'assert' is given in equa-

tion 28 (based on [37]).

$$\begin{aligned}
&\text{Locution: } i \rightarrow j : \text{assert}(p) \\
&\text{Pre-conditions: } (S, p) \in \underline{S}(\Sigma_i \cup CS_j) \\
&\text{Post-conditions: } CS_{i,t+1} = CS_{i,t} \cup \{p\}
\end{aligned} \tag{28}$$

It symbolises that i can make the assertion p to agent j when the argument for p including its support S can be drawn from the set of acceptable arguments \underline{S} of all arguments in the knowledge base of i (Σ_i) combined with the already publicly uttered arguments of j (CS_j). Afterwards, the publicly uttered arguments of i are updated with p for the next iteration (Post-condition). Based on this, Sklar et al. define a lie as the 'assert' speech act with the following pre- and postconditions in equation 29.

$$\begin{aligned}
&\text{Locution: } i \rightarrow j : \text{assert}(p) \\
&\text{Pre-conditions: } (S, \neg p) \in \underline{S}(\Sigma_i \cup CS_j) \text{ AND} \\
&\quad (S', p) \in \underline{S}(\Sigma_i \cup CS_j \cup J_i) \\
&\text{Post-conditions: } CS_{i,t+1} = CS_{i,t} \cup \{p\}
\end{aligned} \tag{29}$$

The modified pre-conditions show that the contrary argument $\neg p$ is possible given only the knowledge base of i and the publicly uttered commitments of j . However, by adding additional arguments J_i which are not in the knowledge base of the agent i originally, he can argue for p .

This definition clearly contains the statement condition (by using the 'assert' speech act), it contains an addressee (j) and the believe-false condition ($(S, \neg p) \in \underline{S}(\Sigma_i \cup CS_j)$). However, the intention to deceive is missing.

The authors indicate that the additionally used knowledge J_i needs to be remembered and maintained for each communication partner to ensure that the agent does not contradict itself. Furthermore, they think about possible applications, e.g. that lying might be an easier option when two different arguments are possible, one truthful but very complicated argument and one easy but dishonest argument. Moreover, they imagine the application area of negotiation where agents want to lie about their personal value of goods.

All in all, speech acts allow for a higher level of abstraction and at the same time for good definitions as well. The problem that occurred in the paper of Sklar et al. and which might be a general problem for this approach is that definitions are speaker-focused and do not account for intentions to deceive or even deception which is solely focused on the addressee.

High-level approaches Still another level of abstraction can be achieved by constructing abstract argumentation frameworks as introduced by Dung in [14]. They have the advantage of providing an easy problem understanding as they are able to depict the situation graphically. On the other hand, we will see that the high level of abstraction does not allow for unambiguous, clear definitions.

The abstract argumentation framework $\langle \mathcal{A}, \rightarrow \rangle$ contains a set of arguments \mathcal{A} and a defeat relation \rightarrow between them. If one argument defeats

another, it contradicts either the original argument itself or one of its supportive propositions. These argumentation frameworks are usually used to easily construct the set of arguments that should be accepted, in the sense that this acceptable set is consistent and does not contradict itself. Several measures can be applied for this purpose (e.g. a grounded semantic containing only the minimal set of arguments that need to be accepted in any case). When applying this approach to agent communication, each agent might propose its own set of acceptable arguments determined by their own knowledge. All publicly announced arguments form a new framework where the overall accepted arguments can be determined. Agents might have preferences over the finally accepted arguments, which in turn are useful when analysing the argumentation game-theoretically. Based on the argumentation frameworks, it is possible to define lies or especially withholding information as agents might want to improve their utility by influencing the acceptable arguments by adding additional arguments or hiding some arguments to break defeat chains.

The paper we will examine here is by I. Rahwan et al. [42]. An almost identical approach has also been described in their previous paper in [40]. Formally, each agent i in this paper has a true type \mathcal{A}_i reflecting the true set of acceptable arguments given the agent's knowledge and a semantic. However, the publicly revealed type \mathcal{A}_i^* might be different from \mathcal{A}_i . (Thoroughly honest agents always have $\mathcal{A}_i^* = \mathcal{A}_i$). Additionally, each agent has a utility function $u_i(\mathcal{A}_1^*, \mathcal{A}_2^*, \dots, \mathcal{A}_i^*, \dots)$ for the possible final outcomes determined by the revealed types of all participants.

Based on this, the authors define their dishonest notions as in equations 30 and 31.

$$\text{Lying: } \mathcal{A}_i^* \not\subseteq \mathcal{A}_i \quad (30)$$

$$\begin{aligned} &\text{With. Inf.: } \mathcal{A}_i^* \subset \mathcal{A}_i \\ \text{AND } &u_i(\mathcal{A}_1^*, \mathcal{A}_2^*, \dots, \mathcal{A}_i^*, \dots) > u_i(\mathcal{A}_1^*, \mathcal{A}_2^*, \dots, \mathcal{A}_i, \dots) \end{aligned} \quad (31)$$

It should be noted that the second condition for withholding information (the comparison between the utility functions) was not given formally, but in words.

Both definitions implicitly include the statement and the addressee condition, since the evaluation of the utility function is based on the preceding communicative act of each agent. One might also argue that the intention to deceive is included as well, because the other agents need to accept the wrongly revealed arguments which presumably is the intention of the agent. However, the believe-false condition for lying is not fulfilled as \mathcal{A}_i^* only needs to contain additional arguments about which nothing is known. Consequently, the lying definition might cover bullshit and lying possibly even paired with withholding information. The same problem holds for the withholding information definition in equation 31 as the higher utility does not suffice to show that the agent is inducing knowledge it does not believe. The definition would fit better to half-truths, as a true part of the knowledge is revealed with the intention to increase one's utility.

In the rest of the paper, the authors show that agents might have an incentive to lie as a Nash equilibrium for the argumentation game does not necessarily lead to honest behaviour. They conclude by proposing that this should be avoided by using restrictive rules. Another nice example for this game-theoretical analysis is shown in Rahwan et. al.'s subsequent paper [41]. Also, other authors like M. Caminada try to use abstract argumentation frameworks to define lies and deception, but without giving formal definitions [7].

Altogether, this high level approach showed to be highly ambiguous when it comes to check the philosophical conditions. However, the practical implications for discussions and argumentation can be derived and investigated more easily due to the high abstraction.

Summary Summarising the other approaches based on different level of agent communication, we can see some common problems. Embedding definitions in a context often either requires to assume implicitly some conditions or to leave out some conditions completely. Moreover, the definitions are fixed on the speaker, which is why no formal definition for deception was given. On the positive side, we can note the high application-orientation of all approaches which has been proved by applying the respective definitions to sensible examples. Moreover, higher levels of abstraction allow for high-level reasoning which might be impossible to do with a low-level logic-only approach. Given that agent technology will develop further, this area of embedding dishonesty in existing frameworks will surely become more important than the clear-cut modal logic definitions.

5 Open issues and conclusion

In this review, we examined formal-logical definitions of lies, deception and other notions. Using the philosophical literature, we extracted guidelines that helped us to check whether formal definitions are correct and complete. Primarily, we looked at purely logical definitions and found some complying with all or almost all necessary conditions. We also assessed a few examples of other approaches which were formally defined but not based on pure logic. These were not as clear-cut and obviously correct, but had the great advantage of being embedded in an application environment. On the other hand, the modal logic definitions were accurate and concise (at least some), but not trouble-free applicable to applications.

In the remaining part we will look at the open issues that arise from this basis. At first, we start by listing some issues identified by the authors of the various papers, before adding other additional issues.

One question that seems to be unanswered, as almost without exception all authors mentioned in chapter 4 referred to it, is how to include lying and deception successfully in agent communication and argumentation. Although some tried, as shown in chapter 4, there is yet no widely accepted solution that would allow to use the notion of lies or deception in practice.

Strongly connected is the question of how to detect lies or deception in communication (e.g. pointed out by Sakama in [49]). There are papers considering this question, e.g. in [26], but not in a formal way which would allow to apply some detection algorithms.

A probably even more useful question is how to prevent agents from lying or trying to deceive [47,42]. This connects to the question of how to design agent communication mechanisms that provide no incentive to lie, thereby enforcing honesty. Similarly, such questions have been considered in an informal way [2] but not applied formally.

Other open questions that have been identified by others are for example the influence of lies or bullshit arguments in discussions and if this could lead to some form of collective irrationality [7]. Or which computational complexity does reasoning about a strategy to lie or deceive actually have and if this is a hindrance to adopt such strategies in practice [42].

An issue which has not been addressed by any of the authors is the trade-off between simplicity/applicability and correctness. All authors using modal logic tried to use as complicated operators as necessary to define lies as correct as possible, whereas authors who included the definition in agent design used operators as simple as possible sacrificing correctness and conciseness. Possible solutions might need to adopt simpler operators which are usable in actual agents to define the notions as correctly as modal operators allowed. Besides simplification, unification of different ideas to one definition instead of many different ones might also help to apply the result in practice.

A second point which has not been addressed widely is how to react on detected lies or attempted deception. Although this step naturally succeeds the still open detection question, it is highly important as it alters the way agents communicate with another. They could introduce some kind of punishment or even ignore the liar in future encounters. The kind of expected punishment also influences the decision to lie as it changes the expected utility.

All in all, we can see that there are a number of open issues and possible improvements to be addressed. One can hope that these issues can be resolved to have, one day, independent self-acting agents confidently using lies, bullshit, deception or other notions avoiding detection, while at the same time being able to detect dishonesties of others in an environment that encourages honest behaviour and has clear rules on how to react to dishonest agents.

References

1. ADLER, J. E. Lying, deceiving, or falsely implicating. *The Journal of Philosophy* 94, 9 (1997), 435–452.
2. AHERN, C., AND CLARK, R. Mechanisms for enforcing honest signaling. Sep 2012.
3. BARBER, K., FULLAM, K., AND KIM, J. Challenges for trust, fraud and deception research in multi-agent systems. In *Trust, Reputation, and Security: Theories and Practice*, R. Falcone, S. Barber, L. Korba,

- and M. Singh, Eds., vol. 2631 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2003, pp. 167–174.
4. BARNES, A. *Seeing through self-deception*. Cambridge University Press, 2007.
 5. BRATMAN, M. *Intention, Plans, and Practical Reason*. Center for the Study of Language and Information, 1987.
 6. CAMBRIDGE UNIVERSITY PRESS. *Cambridge Advanced Learner's Dictionary*, third edition ed. Cambridge University Press, 2008.
 7. CAMINADA, M. W. Truth, lies and bullshit; distinguishing classes of dishonesty. In *Social Simulation workshop at the International Joint Conference on Artificial Intelligence* (2009), pp. 39–50.
 8. CARSON, T. L. *Lying and Deception: Theory and Practice*. OUP Oxford, 2010.
 9. CHELLAS, B. F. *Modal Logic: An Introduction*. Cambridge University Press, 1980.
 10. CHISHOLM, R. M., AND FEEHAN, T. D. The intent to deceive. *The Journal of Philosophy* 74, 3 (1977), 143–159.
 11. COLOMBETTI, M. A modal logic of intentional communication. *Mathematical Social Sciences* 38 (1999), 171–196.
 12. DE ROSIS, F., CAROFIGLIO, V., GRASSANO, G., AND CASTELFRANCHI, C. Can computers deliberately deceive? a simulation tool and its application to turing's imitation game. *Computational Intelligence* 19, 3 (2003), 235–263.
 13. DEMOLOMBE, R. Reasoning about trust: A formal logical framework. In *Trust Management*, C. Jensen, S. Poslad, and T. Dimitrakos, Eds., vol. 2995 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 291–303.
 14. DUNG, P. M. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77, 2 (1995), 321–357.
 15. FALLIS, D. Lying and deception. *Philosopher's Imprint* 10, 11 (Nov 2010), 1–20.
 16. FIROZABADI, B., TAN, Y., AND LEE, R. *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science*. IOS Press, Amsterdam, 1998, ch. Formal definitions of fraud, pp. 275–288.
 17. FOUNDATION FOR INTELLIGENT PHYSICAL AGENTS. Fipa acl message structure specification. <http://www.fipa.org/specs/fipa00061/>, Dec 2002.
 18. FRANKFURT, H. G. *On Bullshit*. Princeton University Press, Princeton, 2005.
 19. FULLER, G. Other-deception. *Southwestern Journal of Philosophy* 7, 1 (1976), 21–31.
 20. GARSON, J. Modal logic. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., spring 2013 ed. 2013.
 21. GRANDISON, T., AND SLOMAN, M. A survey of trust in internet applications. *Communications Surveys Tutorials, IEEE* 3, 4 (2000), 2–16.
 22. HUANG, J., AND NICOL, D. A formal-semantics-based calculus of trust. *Internet Computing, IEEE* 14, 5 (2010), 38–46.

23. JONES, A. J. I., AND FIROZABADI, B. S. *Trust and Deception in Virtual Societies*. Springer, 2001, ch. On the characterisation of a trusting agent - aspects of a formal approach, pp. 157–168.
24. JONES, A. J. I., AND PÖRN, I. Ideality, sub-ideality and deontic logic. *Synthese* 65 (1985), 275–290.
25. JONES, A. J. I., AND SERGOT, M. A formal characterisation of institutionalised power. *Logic Journal of IGPL* 4, 3 (1996), 427–443.
26. JONES, J. Using logic programming to detect deception on the basis of actions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3975 LNCS (2006), 662–663.
27. KEMP, K. W., AND SULLIVAN, T. Speaking falsely and telling lies. *Proceedings of the American Catholic Philosophical Association* 67 (1993), 151–170.
28. KUPFER, J. The moral presumption against lying. *The Review of Metaphysics* 36, 1 (1982), 103–126.
29. LINSKY, L. Deception. *Inquiry* 6, 1-4 (1963), 157–169.
30. LUCK, M., MCBURNEY, P., PREIST, C., GUILFOYLE, C., BERGAMASCHI, S., DAVIDSSON, P., DIGNUM, F., EDWARDS, P., KLUSCH, M., KUDENKO, D., MOSS, S., PETTA, P., ROTH, V., SIERRA, C., AND ZAMBONELLI, F. *Agent technology: enabling next generation computing (a roadmap for agent based computing)*. AgentLink/University of Southampton, 2003.
31. MAHON, J. E. A definition of deceiving. *International Journal of Applied Philosophy* 21, 2 (2007), 181–194.
32. MAHON, J. E. The definition of lying and deception. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., fall 2008 ed. 2008.
33. MAHON, J. E. Two definitions of lying. *International Journal of Applied Philosophy* 22, 2 (Fall 2008), 211–230.
34. MEGGLE, G. Logik der täuschung. In *Vorträge des 3. Internationalen Kongresses der Gesellschaft für Analytische Philosophie* (Berlin / New York, 2000), pp. 339–348.
35. O’NEILL, B. A formal system for understanding lies and deceit. Script based on a talk at the Jerusalem Conference on Biblical Economics, Hebrew University, June 2000, April 2003.
36. PAN, Y., CAO, C., AND SUI, Y. A formal system for lies based on speech acts in multi-agent systems. In *Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence* (2007), pp. 228–234.
37. PARSONS, S., WOOLDRIDGE, M., AND AMGOUD, L. On the outcomes of formal inter-agent dialogues. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems* (2003), ACM, pp. 616–623.
38. PRIMORATZ, I. Lying and the ‘methods of ethics’. *International Studies in Philosophy* 16, 3 (1984), 35–57.
39. PÖRN, I. *Action theory and social science : some formal models*. Dordrecht, Holland ; Boston : D. Reidel Pub. Co, 1977.
40. RAHWAN, I., AND LARSON, K. Mechanism design for abstract argumentation. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 2* (Richland,

- SC, 2008), AAMAS '08, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1031–1038.
41. RAHWAN, I., AND LARSON, K. Argumentation and game theory. In *Argumentation in Artificial Intelligence*, G. Simari and I. Rahwan, Eds. Springer, 2009, pp. 321–339.
 42. RAHWAN, I., LARSON, K., AND TOHME, F. Game-theoretic foundations for argumentation. *Journal of ACM* V (2009), 1–38.
 43. ROSENSCHEIN, J. S., AND ZLOTKIN, G. *Rules of encounter: designing conventions for automated negotiation among computers*. MIT press, 1994.
 44. SAKAMA, C. Logical definitions of lying. In *Proceedings of the 14th International Workshop on Trust in Agent Societies (TRUST11)* (Taipei, Taiwan, May 2011).
 45. SAKAMA, C. A formal model of dishonest communication. Slides of a talk given at the Workshop on Formal Models of Communication, Opole, Poland, August 2012.
 46. SAKAMA, C. Learning dishonesty. Slides of a talk given at the 22nd International Conference on Inductive Logic Programming (ILP 2012), Dubrovnik, Croatia, September 2012.
 47. SAKAMA, C., AND CAMINADA, M. The many faces of deception. In *Proceedings of the Thirty Years of Nonmonotonic Reasoning (Non-Mon@30)* (Lexington, KY, USA, October 2010).
 48. SAKAMA, C., CAMINADA, M., AND HERZIG, A. A logical account of lying. In *Logics in Artificial Intelligence*, T. Janhunen and I. Niemelä, Eds., vol. 6341 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 286–299.
 49. SAKAMA, C., SON, T. C., AND PONTELLI, E. A logical formulation for negotiation among dishonest agents. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (2011), pp. 1069–1074.
 50. SANTOS, F., JONES, A., AND CARMO, J. Action concepts for describing organised interaction. In *Proceedings of the Thirtieth Hawaii International Conference on System Sciences* (Jan 1997), vol. 5, pp. 373–382.
 51. SEARLE, J. R. A classification of illocutionary acts. *Language in society* 5, 01 (1976), 1–23.
 52. SIMMONS, M. Recognizing the elements of fraud. www.facilitatedcontrols.com/fraud-investigation/fraudwww.shtml, 1995. Retrieved: 11th February 2013.
 53. SKLAR, E., PARSONS, S., AND DAVIES, M. When is it okay to lie? a simple model of contradiction in agent-based dialogues. In *Argumentation in Multi-Agent Systems*, I. Rahwan, P. Moraitis, and C. Reed, Eds., vol. 3366 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 251–261.
 54. SOLOMON, N. Guidance on certain manipulative and deceptive trading practices. Guidance Note 13-0053 of Investment Industry Regulatory Organization Canada (IIROC), February 2013.
 55. SORENSEN, R. Bald-faced lies! lying without the intent to deceive. *Pacific Philosophical Quarterly* 88, 2 (June 2007), 251–264.

56. STECH, F., HECKMAN, K. E., HILLIARD, P., AND BALLO, J. R. Scientometrics of deception, counter-deception, and deception detection in cyber-space. *PsychNology Journal* 9, 2 (2011), 79–122.
57. STRANDERS, R. Argumentation based decision making for trust in multi-agent systems. *IEEE Transactions on Knowledge and Data Engineering* 6, 1 (2006), 64–71.
58. TURING, A. M. Computing machinery and intelligence. *Mind* LIX, 236 (1950), 433–460.
59. TZOUVARAS, A. Logic of knowledge and utterance and the liar. *Journal of Philosophical Logic* 27, 1 (1998), 85–108.
60. VRIJ, A. *Detecting Lies and Deceit: The psychology of lying and the implications for professional practice*. John Wiley and Sons, Chichester, 2000.
61. WALTON, D. N., AND KRABBE, E. C. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. Suny Press, 1995.
62. WILLIAMS, B. *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press, February 2004.